ESTIMATING THE PARAMETERS OF A TRUNCATED GAMMA DISTRIBUTION

By

Douglas G. Chapman

University of Washington

Technical Report No. 13

September 2, 1953

Contract N8onr-520 Task Order II
Project Number NR-042-038

Laboratory of Statistical Research
Department of Mathematics
University of Washington
Seattle, Washington

Estimating the Parameters of a Truncated Gamma Distribution [1]

Douglas G. Chapman
University of Washington

1. Summary. A table is given to simplify the estimation of the parameters of an incomplete gamma or Type III distribution. A new procedure is also suggested to estimate the parameters of a truncated gamma distribution. This method is also applicable to a number of other truncated distributions, whether the truncation is in the tails or the center of the distribution.

2. Introduction. Several examples have been given recently, employing the incomplete gamma or Type III distribution in fitting rainfall data, e.g. [1,2]. In an animal population study [3], it was found that the migration pattern could be fitted by this type of distribution. Frequently in such migration studies the data will be truncated, i.e. observations will begin after migration has begun or conclude before it has stopped.

The parameters of the gamma distribution are often estimated by the method of moments in such cases [see for example [4] pp. 121, 125] despite the fact that Fisher [5] showed the method to be inefficient. To facilitate solution of the maximum likelihood equations for estimation of the parameters in the untruncated case, a simple table is given.

The estimation of the parameters of a truncated gamma distribution, by the method of moments, has been studied by Cohen [6]. Since the

[1] Research sponsored by the Office of Naval Research.

integral of the probability density cannot be expressed in closed form, even the moment estimates are tedious to obtain; no attempt has been made to evaluate their variances or study their efficiencies. After this paper was completed, a new study of the problem was published by Des Raj [13]. He gives the maximum likelihood equations for a number of cases of truncated and censored samples, mainly however under the assumption that the third standard moment is known. These equations can be solved only by iterative methods. In this paper a new method of estimation of these parameters is introduced which is easier to apply. The asymptotic variance-covariance matrix of the estimates is determined.

3. Estimation with origin known. The density function of the gamma distribution may be written in the form

$$f(x) = \frac{a^b}{\Gamma(b)} e^{-(x-c)}(x-c)^{b-1} \qquad x \geq c$$

$$= 0 \qquad x < c$$

The parameters are frequently transformed so that the distribution is expressed as a function of the mean, variance and skewness. Since the corresponding sample quantities do not efficiently estimate the parameters such a transformation appears to be misleading.

The maximum likelihood equations have been given by Fisher [5], viz.

(1) $$\frac{\partial L}{\partial a} = \frac{b}{a} - (\bar{x} - c) = 0$$

(2) $$\frac{\partial L}{\partial b} = \ln a - \frac{\Gamma'(b)}{\Gamma(b)} + \frac{1}{n}\sum_{i=1}^{n} \ln(x_i - c) = 0$$

(3) $$\frac{\partial L}{\partial c} = a - \frac{b-1}{n}\sum_{i=1}^{n} \frac{1}{(x_i - c)} = 0$$

Since the parameter c determines the region of positive density, that equation (3) gives the maximum likelihood estimate of c must be justified in a slightly different manner than by routine calculus. If $b > 1$ this is easily done; if, however, $b \leq 1$ equation (3) does not give the maximum likelihood of c. In this case f(x) is monotone decreasing for $x \geq c$ and $s = \min_i x_i$ is the maximum likelihood estimate of c.

We consider first the case where the origin is known: so that c may be set equal to zero, without loss of generality and equation (3) drops out. Letting

$$\overline{x}_L = \frac{1}{n} \sum_{i=1}^{n} \ln x_i$$

(1) and (2) yield

$$(4) \qquad \gamma(b) = \ln b - \frac{\Gamma'(b)}{\Gamma(b)} = \ln \overline{x} - \overline{x}_L$$

Since $\frac{\Gamma'(b)}{\Gamma(b)}$, the digamma function, has been tabulated by Pairman [7], it is easy to construct a table of $\gamma(b)$ and solve for b by inverse interpolation. A small tabulation of $\gamma(b)$ is given in Table I; a more complete tabulation is available in mimeographed form from the Laboratory of Statistical Research, University of Washington. There $\gamma(b)$ and its first and second differences are tabulated for b=1 (.01)5, 5(0.1)20; 20(1)100. The table was checked by summing columns in the basic tables and should be correct to one figure in the fifth decimal.

4. Estimation in the truncated case with known origin. The density function is now written

$$(4) \qquad f(x) = K^{-1} e^{-ax} x^{b-1} \qquad 0 \leq x \leq \tau$$

$$= 0 \text{ elsewhere}$$

$$\text{where } K(a,b) = \int_0^{\tau} e^{-ax} x^{b-1} dx$$

The maximum likelihood functions now involve derivatives of K with respect to a and b respectively; a double entry table would be necessary to obtain the maximum likelihood estimates of a and b and even this would involve double inverse interpolation.

In lieu of this another method of estimation is proposed. Let the n observations be grouped by classes $(\xi_i - h_i, \xi_i + h_i)$, $(i = 1,2,\ldots r)$ where $\xi_1 - h_1 = 0$, $\xi_r + h_r = \tau$ and denote by ν_i the number of observations falling in class i, i.e. between $\xi_i - h_i$ and $\xi_i + h_i$.

Define

$$(5) \qquad p_i = K^{-1} \int_{\xi_i - h_i}^{\xi_i + h_i} e^{-ax} x^{b-1} dx \doteq K e^{-a\xi_i} \xi_i^{b-1} (2h_i)$$

$$q_i = \frac{\nu_i}{n}$$

Now

$$(6) \qquad \ln p_i - \ln p_{i+1} = a(\xi_{i+1} - \xi_i) + (b-1) \ln \frac{\xi_i}{\xi_{i+1}} + \ln \frac{h_i}{h_{i+1}}$$

$(i = 1,2,\ldots r-1)$ to the degree of approximation indicated by (5).

The form of equation (6) suggests estimating a and b by a least squares procedure, with q_i replacing p_i. This can be justified as an approximate procedure by the following results. To terms of order $\frac{1}{n}$

$$(7) \qquad \mathcal{E}(\ln q_i) = \ln p_i - \frac{1}{2n} \frac{1-p_i}{p_i}$$

$$(8) \qquad \mathcal{E}(\ln q_i - \ln p_i)^2 = \frac{1}{n}\,\frac{1-p_i}{p_i}$$

$$(9) \qquad \mathcal{E}\left(\ln \frac{q_i}{p_i}\right)\left(\ln \frac{q_j}{p_j}\right) = -\frac{1}{n}$$

These results can be obtained by expanding $\ln(q_i/p_i) = \ln\left(1+\frac{q_i-p_i}{p_i}\right)$ in a Taylor series (assuming that for large n, $\Pr(q_i = 0)$ and $\Pr(q_i > 2p_i)$ may be neglected). It is also necessary to use the fact that $\mathcal{E}(q^r) = 0\left(\frac{1}{n^r}\right)$, a result easily obtainable from the well-known recurrence formula for the central moments of the multinomial distribution, viz.

$$\mu_{r+1} = pq\left(nr\,\mu_{r-1} + \frac{d\mu_r}{dp}\right)$$

From this, in fact, it may be inferred that

$$\mathcal{E}(q^{2r+1}) = 0\left(\frac{1}{n^{r+1}}\right)$$

$$\mathcal{E}(q^{2r}) \quad = 0\left(\frac{1}{n^r}\right) \qquad r \geq 1$$

Moreover, the limiting distribution of the $\ln q_i$ is easily obtained from the following lemma.

Lemma Let $\left\{X_i^{(n)}\right\}$ $(i = 1,2,\ldots,r)$ be a sequence of random variables and μ_i, σ_i $(i=1,2,\ldots,r)$ be constants such that the joint distribution of

$$Y_i^{(n)} = \frac{(X_i^{(n)} - \mu_i)}{\sigma_i}\sqrt{n} \qquad (i = 1,2,\ldots,r)$$

tends to the limiting distribution $F(y_1,y_2,\ldots,y_r)$ as $n \to \infty$.

and let $f(x)$ be of class $C^{(1)}$ in the neighborhood of $(\mu_1, \mu_2, \ldots, \mu_r)$ with $\tau_i = \left(\frac{df}{dx}\right)_{x=\mu_i} \neq 0$

then $$z_i^{(n)} = \frac{\sqrt{n}\,\left[f(X_i^{(n)}) - f(\mu_i)\right]}{\sigma_i \cdot \tau_i} \qquad (i = 1,2,\ldots,r)$$

have the same joint limiting distribution.

This is a consequence of the general theorems on stochastic limit relationships proved by Mann and Wald [8] (cf. their Theorems 3 and 5-a trivial modification is however required since our $f(x)$ is a function of a single real variable while their corresponding $g(x)$ is a function of a vector-valued random variable).

Finally, writing

(10) $$y_i = \ln q_i - \ln q_{i+1} \qquad (i = 1,2,\ldots,r-1)$$

it follows that the y_i are asymptotically multinormal with means

$$a\left(\xi_{i+1} - \xi_i\right) + (b-1)\ln\frac{\xi_i}{\xi_{i+1}} + \ln\frac{h_i}{h_{i+1}}$$

and moment matrix

$$\mathcal{M} = \begin{pmatrix} \frac{1}{n}\left(\frac{1}{p_1}+\frac{1}{p_2}\right), & -\frac{1}{n}\left(\frac{1}{p_2}\right), & 0 & \cdots & 0 \\ -\frac{1}{n}\left(\frac{1}{p_2}\right) & , \frac{1}{n}\left(\frac{1}{p_2}+\frac{1}{p_3}\right) & , -\frac{1}{n}\left(\frac{1}{p_3}\right) & \cdots & 0 \\ 0 & , -\frac{1}{n}\left(\frac{1}{p_3}\right) & , -\frac{1}{n}\left(\frac{1}{p_3}+\frac{1}{p_4}\right) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & , 0 & , 0, & \ldots, & \frac{1}{n}\left(\frac{1}{p_{r-1}}+\frac{1}{p_r}\right) \end{pmatrix}$$

Asymptotically efficient estimators of a and b are found by minimizing the quadratic form

$$\overrightarrow{y - \mathcal{E}(y)}' \; \mathcal{M}^{-1} \; \overrightarrow{y - \mathcal{E}(y)}$$

where the vector $\vec{y} = (y_1, y_2, \ldots y_{r-1})$

Since the true values of the p_i are not known, it is necessary to replace the p_i in $\mathcal{M}$ by their estimates, the q_i. Introducing the notation

(11) $$w_i = y_i - \ln h_i + \ln h_{i+1}$$

(12) $$u_i = \xi_{i+1} - \xi_i$$

(13) $$v_i = \ln \xi_i - \ln \xi_{i+1}$$

the equations for a and b are

(14) $$a\left(\sum_i \sum_j m_o^{ij} u_i u_j\right) + b\left(\sum_i \sum_j m_o^{ij} u_i v_j\right) = \sum_i \sum_j m_o^{ij} u_i w_j$$

(15) $$a\left(\sum_i \sum_j m_o^{ij} u_i v_j\right) + b\left(\sum_i \sum_j m_o^{ij} v_i v_j\right) = \sum_i \sum_j m_o^{ij} v_i w_j$$

m_o^{ij} denoting the elements of $\mathcal{M}_o^{-1}$ ($\mathcal{M}^{-1}$ with p's replaced by q's)

The solutions of these are

(16) $$\hat{a} = \frac{1}{\Delta}\left[(\vec{v}' \mathcal{M}_o^{-1} \vec{v})(\vec{u}' \mathcal{M}_o^{-1} \vec{w}) - (\vec{u}' \mathcal{M}_o^{-1} \vec{v})(\vec{v}' \mathcal{M}_o^{-1} \vec{w})\right.$$

(17) $$\hat{b} = \frac{1}{\Delta}\left[(\vec{u}' \mathcal{M}_o^{-1} \vec{u})(\vec{v} \mathcal{M}_o^{-1} \vec{w}) - (\vec{u} \mathcal{M}_o^{-1} \vec{v})(\vec{u}' \mathcal{M}_o^{-1} \vec{w})\right.$$

where $$\Delta = (\vec{u}' \mathcal{M}_o^{-1} \vec{u})(\vec{v}' \mathcal{M}_o^{-1} \vec{v}) - (\vec{u}' \mathcal{M}_o^{-1} \vec{v})^2$$

and the covariance matrix of $(\hat{a}, \hat{b})$ is

$$\begin{pmatrix} \frac{1}{\Delta}(\vec{v}' \mathcal{M}^{-1} v) & -\frac{1}{\Delta}(\vec{u}' \mathcal{M}^{-1} \vec{v}) \\ -\frac{1}{\Delta}(u' \mathcal{M}^{-1} \vec{v}) & \frac{1}{\Delta}(u' \mathcal{M}^{-1} \vec{u}) \end{pmatrix}$$

The estimates $\hat{a}$ and $\hat{b}$ are found by direct simple routine calculations except for the determination of $\mathcal{M}^{-1}$ from $\mathcal{M}$. This may be a tedious process unless r is small. However, if all p_i are equal to $\frac{1}{r}$, then

$$\frac{r^2}{n}\mathcal{M}^{-1} = \begin{pmatrix} r-1, & r-2, & r-3, & \dots & 3, & 2, & 1 \\ r-2, & 2(r-2), & 2(r-3), & \dots & 6, & 4, & 2 \\ r-3, & 2(r-3), & 3(r-3), & \dots & 9, & 6, & 3 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 3, & 6, & 9, & \dots & 3(r-3), & 2(r-3), & r-3 \\ 2, & 4, & 6, & \dots & 2(r-3), & 2(r-2), & r-2 \\ 1, & 2, & 3, & \dots & (r-3), & r-2, & r-1 \end{pmatrix}$$

This is easily verified by direct multiplication. This suggests that where possible the ξ_{ri} should be chosen so that the q_i (and thus approximately, the p_i) are equal. The device is analogous to that suggested by Gumbel [9] and by Mann and Wald [10] in applying the χ^2 "goodness-of-fit" test.

If this is not possible, less efficient estimates can be obtained by utilising only the odd (or even) w_i's. The odd w_i's are mutually independent among themselves and consequently $\mathcal{M}$ and $\mathcal{M}^{-1}$ reduce to diagonal matrices.

5. Estimation with unknown origin. If the parameter c, the origin, is unknown, then the estimation problem is more difficult whether or not the

distribution is truncated. Iterative methods are of course possible in solving (1) (2) and (3) with the aid of Table I, i.e. for the untruncated case. In the truncated case this method is too tedious to have much practical value.

If, in the truncated case, there is available supplementary information so that the restriction $0 < c << \xi_1$ may be utilized, then a procedure similar to that outlined above may be followed. In this case

$$(18)\quad \ln p_i - \ln p_{i+1} = a(\xi_{i+1} - \xi_i) + (b-1) \ln \frac{\xi_i - c}{\xi_{i+1} - c} + \ln \frac{h_i}{h_{i+1}}$$

$$(i = 1,2,\ldots r-1)$$

again to the degree of approximation indicated by (5). With the restriction noted above, it is adequate to write

$$(19)\quad \ln p_i - \ln p_{i+1} = a(\xi_{i+1} - \xi_i) + (b-1) \ln \frac{\xi_i}{\xi_{i+1}} + (b-1)\, c \left(\frac{1}{\xi_{i+1}} - \frac{1}{\xi_i} \right) + \ln \frac{h_i}{h_{i+1}}$$

Defining y_i, w_i as above least squares estimates of a, b and c may be found in an exactly analogous procedure to that of Section 4.

6. Conclusion. The method used to estimate the parameters a and b in Section 4 may also be applied if the sample is drawn from a doubly truncated gamma distribution, from a singly or doubly truncated normal distribution or from a beta distribution with known range, either truncated or not. Methods of obtaining the maximum likelihood estimates of the parameters of a truncated normal distribution are, of course, well known,

and extensive tabulations have been made to facilitate the determination of such solutions (e.g., compare particularly Hald [11]).

The method outlined above would also be useful in estimating the parameters of the normal curve where there are systematic gaps in the observations. This may occur particularly in time distributions—an example may be found in [12]. For distributions with finite but unknown range, however, the method does not appear to be satisfactory.

References.

1. Report on Bonneville Power Administration Cloud-Seeding Operations, U. S. Dept. of Interior, July 1952, particularly pp. 33 ff.

2. T. A. Jeeves, L. M. LeCam, J. Neyman, and E. L. Scott, "Problem of documentary evidence of the effectiveness of cloudseeding". Unpublished report.

3. Karl W. Kenyon, Victor B. Scheffer, Douglas G. Chapman, "A Population study of the Alaska fur seal herd". To be published.

4. J. F. Kenney and E. S. Keeping, Mathematics of Statistics, Part Two. D. Van Nostrand, Inc., New York (1951).

5. R. A. Fisher, "On the mathematical foundations of theoretical statistics", Phil. Trans. Royal Soc. of London, Series A 222 (1922), pp. 309-368.

6. A. C. Cohen, "Estimating parameters of Type III populations from truncated samples", Jour. Amer. Stat. Assoc. 45 (1950) pp. 411-423.

7. Eleanor Pairman, Tables of the Digamma and Trigamma functions. Tracts for Computers I (edited by Karl Pearson). Cambridge University Press, London, 1919.

8. H. B. Mann and A. Wald, "On Stochastic Limit and Order Relationships", Ann. Math. Stat. 14 (1943), pp. 217-226.

9. E. J. Gumbel, "On the reliability of the classical Chi-Square test", Ann. Math. Stat. 14 (1943), pp. 253-263.

10. H. B. Mann and A. Wald, "On the choice of the number of intervals in the application of the Chi-Square test". Ann. Math. Stat. 13 (1942) pp. 306-317.

11. A. Hald, "Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point", Skand. Aktuarietidskr. 32 (1949), pp. 119-134.

12. "Interim report on submerged orifice research, powerhouse fish collection system, Bonneville Dam Investigations August - September 1952", Corps of Engineers, Department of the Army, 1953.

13. Des Raj, "Estimation of the parameters of Type III populations from truncated samples", Jour. Amer. Stat. Assoc. 48 (1953) pp. 336-349.

TABLE I

b	$\gamma(b)$	Δ
1.0	.57722	
		.05816
1.1	.51906	
		.04770
1.2	.47136	
		.03980
1.3	.43156	
		.03370
1.4	.39786	
		.02888
1.5	.36898	
		.02502
1.6	.34396	
		.02188
1.7	.32208	
		.01928
1.8	.30280	
		.01713
1.9	.28567	
		.01531
2.0	.27036	
		.01376
2.1	.25660	
		.01244
2.2	.24416	
		.01129
2.3	.23287	
		.01030
2.4	.22257	
		.00944
2.5	.21313	
		.00867
2.6	.20446	
		.00799
2.7	.19647	
		.00740
2.8	.18907	
		.00686
2.9	.18221	
		.00638
3.0	.17583	
		.00595
3.1	.16988	
		.00557
3.2	.16431	
		.00521
3.3	.15910	
		.00489

3.4	.15421	.00460
3.5	.14961	.00434
3.6	.14527	.00409
3.7	.14118	.00387
3.8	.13731	.00366
3.9	.13365	.00347
4.0	.13018	.00330
4.1	.12688	.00314
4.2	.12374	.00298
4.3	.12076	.00284
4.4	.11792	.00271
4.5	.11521	.00259
4.6	.11262	.00248
4.7	.11014	.00237
4.8	.10777	.00227
4.9	.10550	.00218
5.0	.10332	.009665
5.5	.093655	.008013
6.0	.085642	.006751
6.5	.078891	.005765
7.0	.073126	.004980
7.5	.068146	.004346
8.0	.063800	.003825
8.5	.059975	.003392
9.0	.056583	.003029
9.5	.053554	.002722

10.0	.050832	.004689
11.0	.046143	.003898
12.0	.042245	.003291
13.0	.038954	.002815
14.0	.036139	.002435
15.0	.033704	.002129
16.0	.031575	.001875
17.0	.029700	.001665
18.0	.028035	.001488
19.0	.026547	.001339
20.0	.025208	